

Errors in Machine Translation from Croatian to English: a Corpus-based Approach

Zagorec, Dario

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka, Faculty of Humanities and Social Sciences / Sveučilište u Rijeci, Filozofski fakultet**

Permanent link / Trajna poveznica: <https://um.nsk.hr/um:nbn:hr:186:395434>

Rights / Prava: [In copyright](#) / [Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2024-06-16**



Repository / Repozitorij:

[Repository of the University of Rijeka, Faculty of Humanities and Social Sciences - FHSSRI Repository](#)



UNIVERSITY OF RIJEKA
FACULTY OF HUMANITIES AND SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

Dario Zagorec

**ERRORS IN MACHINE TRANSLATION FROM CROATIAN TO
ENGLISH: A CORPUS-BASED APPROACH**

Submitted in Partial Fulfillment of the Requirements for the M.A. in English
Language and Literature and Pedagogy at the University of Rijeka

Supervisor:
Mirjana Borucinsky, Ph. D.

Rijeka, August 2020

Abstract

A once heavily flawed method of translation, machine translation has improved and continues to improve every day. However, with the plethora of errors that machine translation systems make, machine translation is undeniably inferior to human translation. This paper analyzes a corpus of texts translated by Google Translate for different types of errors. The results show that Google Translate committed an abundance of lexical, morphosyntactic and stylistic errors, highlighting the limitations of machine translation at this time.

Key words: translation, Croatian, English, corpus, machine translation, error

Table of Contents

1. INTRODUCTION	1
2. HISTORY OF MACHINE TRANSLATION.....	2
3. MODERN MACHINE TRANSLATION.....	9
4. ADVANTAGES AND DISADVANTAGES OF MACHINE TRANSLATION	12
5. COMPUTER-ASSISTED TRANSLATION.....	13
6. METHOD.....	14
7. ANALYSIS.....	15
7.1. Statistics	15
7.2. Quantitative analysis	17
7.3. Qualitative analysis	20
7.3.1. Untranslated words	22
7.3.2. Omitted words.....	23
7.3.3. Inserted words	26
7.3.4. Lexical errors	27
7.3.5. Spelling errors.....	31
7.3.6. Morphosyntactic errors.....	33
7.3.7. Stylistic errors	40
7.3.8. Punctuation errors.....	42
7.3.9. Inconsistencies	46
8. DISCUSSION.....	49
9. CONCLUSION.....	54
REFERENCES	56
WEB SOURCES.....	58

1. Introduction

The need for translation is as old as language itself. The first significant translation of the Western world is considered to be the translation of the Hebrew Bible into Greek. Today, translation is field of work that includes translating various literary, commercial, legal, technical, judicial and medical texts, among others. As translation has changed throughout history, so too have our approaches to it.

With the ever-increasing need for information exchange and long-distance communication in the modern world, the demand for language translation is greater than ever before. In an effort to lighten the ever-increasing load now placed on translators, modern technology has offered a seemingly perfect solution: machine translation. If a computer could do the job of a translator just as well if not better than a human and in a fraction of the time, the challenge, tediousness and time intensive work that characterize translation would be turned into child's play. Of course, the reality of the situation is much different. The state of machine translation at the moment is such that most machine-translated texts require a large amount of reworking by a human editor, as the translated texts are frequently riddled with errors which make the texts erroneously confusing and often times humorous. They do, however, offer a gist of what the text is about.

In this study, I will analyze the errors made by a computer in translating a corpus of ten texts from Croatian to English with the goal of uncovering the flaws of machine-translated texts from Croatian to English. I hope that by doing so I will contribute to the better understanding of machine translation for the Croatian-English language pair.

2. History of machine translation

It is commonly believed that the idea of machine translation originated in the 17th century, when the idea of using mechanical dictionaries to overcome language barriers was first introduced. However, the actual origins of machine translation stretch as far back as the 9th century, when the Arabs discovered cryptanalysis during the Islamic Golden Age. DuPont (2018) argues that “cryptanalysis is connected to machine translation in its methods, ontologies and histories.” The most significant name in the field of cryptanalysis during this time is Al-Kindi, who managed to “derive a knowledge of patterns and features within and across languages” through his study of linguistics. Al-Kindi discovered that the ratio of vowels to consonants varies from language to language and he developed three key techniques of cryptanalysis – letter frequencies, a technique with which he was able to compare statistical probabilities across corpora; the second technique with which he analyzed morphology and word derivations to describe “combinable” and “non-combinable” letters; and the third technique, called the “probable word” method, which he used to prove that certain words or phrases could be used as “guides” through the cryptogram. (DuPont, 2018)

In the 16th and 17th century, Western scientists developed techniques of encryption and decryption which differed from the cryptanalysis-based methods introduced by Al-Kindi in order to manipulate texts in the effort to “discover or invent universal, perfect and philosophical languages.” These scientists are known as “universal language planners.” Most of these universal language planners used techniques of encryption and decryption, methods which are quite rudimentary compared to the techniques of Al-Kindi, while some

of them used techniques and machines drawn from the fields of cryptology and cryptanalysis, like Johannes Trithemius, Athanasius Kircher, Francis Lodwick and Cave Beck. These universal language planners marked the western origins of early machine translation. (DuPont, 2018)

As impressive as all of these achievements were, the various universal languages proposed and the techniques that they used up to this point in history were little more than glossaries for other languages and proposals for an interlingua to be used instead of natural languages. It was not until 1924 that the first proposed typewriter-translator was proposed – the first true example of machine translation which would probably have worked as a mechanical dictionary. However, the first true and generally accepted beginning of machine translation was in 1933, when Georges B Artsrouni in France and Petr Trojanskij in Russia simultaneously applied for patents for their respective translation devices. These two devices were, at least in their methods, true precursors to modern machine translation (Hutchins, 1995: 434.)

Georges Artsrouni's device was a multipurpose "mechanical brain" capable of automatically producing stored information such as banking statements, telephone directories, railway timetables etc. It was also claimed to be capable of translation, as it was a device especially suited for cryptography. As a translation machine, it had several downsides: the machine, like its inventor who was not a linguist, "had no awareness of problems of polysemy, idioms, or syntactic ambiguity" (Hutchins, 2002: 1-4), but was instead meant for quick, rough translations of simple messages. It worked by crudely substituting words from the source to the target language. It was also slow, had a limited "memory" capacity and compiling the dictionary was expensive.

Unlike Artsrouni's device, Trojanskij's device was not designed to be multipurpose; it was a device fully intended to be used for translation and it was much more complex in its intended use. It was designed as a desk on which was a belt which acted as a dictionary containing entries in six languages. Trojanskij's device made use of "logical parsing symbols" taken from Esperanto, which represented "universal" grammatical relationships, making it possible to translate to and from any language. Trojanskij's device was intended to be used by multiple people: first, an editor who is familiar with the source language would analyze the words and convert them into base forms according to their syntactic functions, then the machine would organize these base forms into their equivalent forms in the target language and a final editor who knew both languages to "extract the meaning of the translation, to choose synonyms, to polish the unevenness, i.e. to do general literary finishing." (Hutchins, 2002: 4-11)

The next great step in the development of machine translation was in 1949, when Warren Weaver of the Rockefeller Foundation outlined his idea of using computers for translation, which was still based on the word-to-word approach helped along by editors. In 1951 the first full-time researcher into machine translation, Yehoshua Bar-Hillel was appointed and in 1954 a public demonstration of machine translation from Russian into English was performed, which spurred interest into further research. What followed was "the decade of high expectation and disillusion" from 1956 to 1966, a time when many US researchers and institutions focused on developing machine translation, usually to and from Russian for reasons of a political nature. Namely, the power struggle between the United States and Russia during the cold war spurred the need of the two sides to quickly and easily translate documents and technical papers of the other side. This research

included the work of Erwin Reifler at the University of Washington, which epitomized the word-for-word approach¹ and which could only be used to translate written and non-literary texts, the work of Gilbert King at the IBM Corporation, which installed a system of translation for the US Air Force, the work performed at the RAND Corporation, which was corpus-based and utilized statistical analyses, the work of Leon Dostert at Georgetown University, whose peers offered a variety of approaches and methods, the work of Paul Garvin at the Ramo-Wooldridge Corporation and more. Naturally, the Soviet Union performed research into machine translation at this time as well at the Institute of Precision Mechanics, the Steklov Mathematical Institute and at Leningrad University. Other nations around the world also participated but had little long-term influence. In fact, even American efforts bore little fruit: the semantic barriers that machine translation struggled to overcome were highlighted by Bar-Hillel in 1960, and the ALPAC (Automatic Language Processing Advisory Committee) report that followed in 1964 came to the said that “machine translation was slower, less accurate and twice as expensive as human translation and that there is no immediate or predictable prospect of useful machine translation.” (Hutchins, 1995: 9) The ALPAC report was unreasonably critical of the fact that the automatic translations needed to be post-edited by a human and is considered to be shortsighted and biased. (Hutchins, 1995: 4-9)

¹ Word-for-word translation is typical of the Direct Machine Translation (DMT) approach, the oldest method of machine translation, which is characterized by direct translation at the word level. It included frequent mistranslations at the lexical level and inappropriate syntax structure (Okpor, 2014).

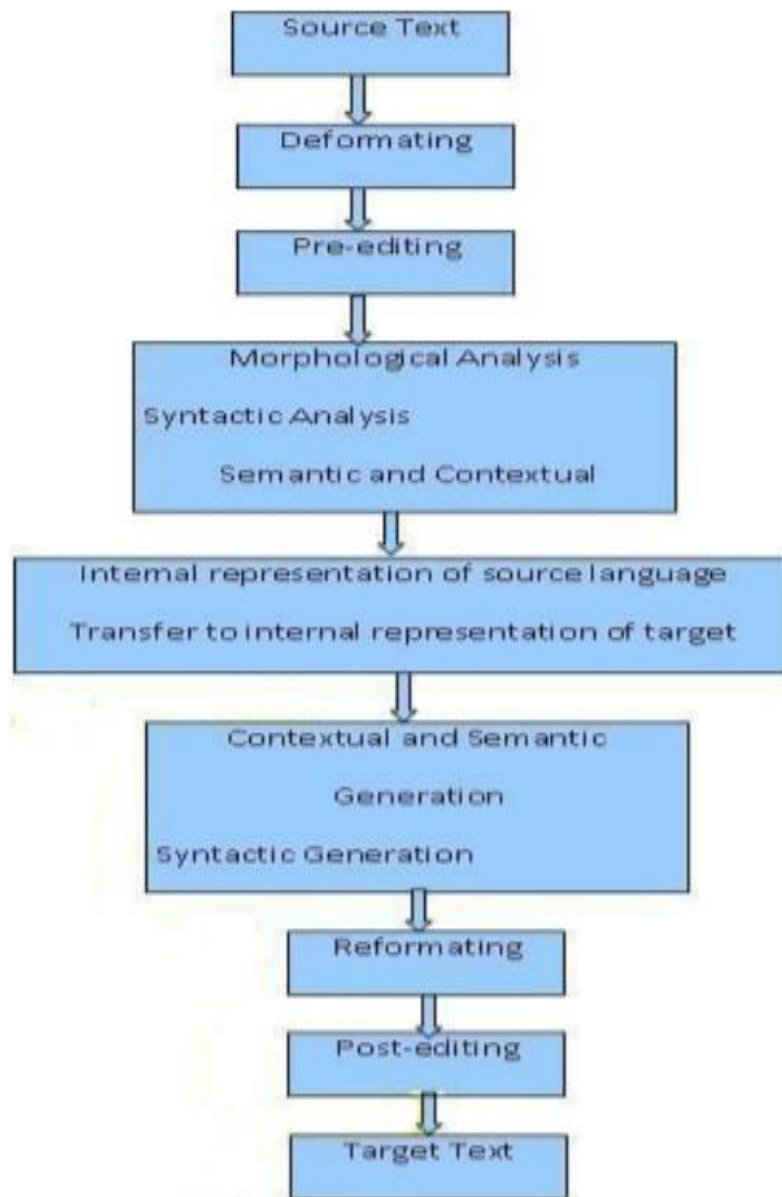


Fig. 1. The typical machine translation process. Source: Okpor, 2014.

Nevertheless, its influence was great. The scalding criticism of the ALPAC report led to a massive decrease in interest in machine translation. What followed was “the quiet decade” from 1967 to 1976, a period of time when research into machine translation in the United States was in hiatus. Research continued in other nations, such as Canada,

where the TAUM project (*Traduction Automatique de l'Université de Montréal*) successfully created a computational metalanguage for manipulating linguistic strings and the Météo system for translating weather forecasts. The LRC at Texas and the CETA (*Centre d'Etudes pour la Traduction Automatique*), as well as the research of Mel'chuk in the Soviet Union focused on an interlingua system of machine translation but by the 1970s that approach started to seem too ambitious and interest in it dropped (Hutchins, 1995: 9-10.)

In the 1980s many new machine translation systems were beginning to appear: TITUS, a multilingual system for translating abstracts written in a controlled language, CULT (Chinese University of Hong Kong), for translating mathematics from Chinese into English, the Logos Corporation, which created a system which specialized in translating aircraft manuals and Systran, the most successful of them all even to this day, which was used by the US Air Force and the European Communities, among others. Worthy of note is also the METAL German-English system which was the most sophisticated commercially available system during the 1980s. These systems were all rule-based; they performed translations based on rules for syntactic analysis, lexical rules, rules for lexical transfer, for syntactic generation, morphology etc. But starting in 1989 example-based and corpus-based methods started to appear, methods which used no syntactic or semantic rules in the analysis of texts or in the selection of lexical equivalents; only statistical methods were applied. The most influential development into corpus-based translation was the Candide project at IBM, which produced remarkably good translations compared to those of rule-based approaches: almost half of the sentences it translated perfectly matched the translations in the corpus or expressed the same thing in slightly

different words. Other projects still focused on rule-based approaches to translation, both in transfer and interlingua systems, like Eurolang, the LMT project run by Michael McCord, the ULTRA system at the New Mexico State University and the Pangloss project, a collaborative project which includes experts from the universities of Southern California, New Mexico State and Carnegie Mellon. In the 1980s a trend towards the adoption of “unification” and “constraint-based” formalism led to a simplification of the rules of analysis, transformation and generation of machine translations. (Hutchins, 1995: 11-20)

In the 1990s, various groups experimented with “dialogue-based machine translation” systems, systems where the text to be translated is composed or written collaboratively between a person and a machine, others experimented with combining speech recognition with translation in order to create spoken language translation. In respect to the development of machine translation, the 1990s are also characterized by the globalization of machine translation research. Systems like Systran, Logos, METAL and ATLAS continued to be used, now with a focus on large scale production of translations which primarily includes technical documentation. Technical documentation contains formulaic language and sequences which can simply be substituted 1:1, which is easy for translations systems to translate. Focus has shifted from attempting to replace humans with machines for translation to “special domain-restricted mass-volume systems and on systems for non-translators – areas where professional translators have not been active.” (Hutchins, 1995: 22) Many translation systems also became available for personal computers. (Hutchins, 1995: 20-22)

3. Modern machine translation

The shift from rule-based to Statistical Machine Translation (SMT) was significant for the development of machine translation. SMT methods, which are generally phrase-based, “have been the dominant translation paradigm for decades.” (Wu et. al., 2016: 2) They have been used extensively since 2002, but in recent years a new type of machine translation was developed: Neural Machine Translation (NMT), (Srivastava et al., 2018: 1.) NMT is, in short, “the use of neural network models to learn a statistical model for machine translation” (Brownlee, 2017.) According to Wu et. al. (2016: 1), NMT has “the potential to overcome many of the weaknesses of conventional phrase-based translation systems.” It works by learning the mapping from input text to associated output text in an end-to-end fashion. It is made up of two recurrent neural networks (RNNs), one which consumes the input text sequence and one which generates translated output text (Wu et. al., 2016: 1.) These RNNs allow for variable length input and output sequences because they are organized into an encoder-decoder architecture. The model can encode the source text in a context vector – an internal fixed-length representation which could, in principle, be decoded i.e. translated into different languages. This type of approach means that one system can be programmed directly on both the source and the target text, removing the need for the complex specialized systems used in SMT. (Brownlee, 2017)

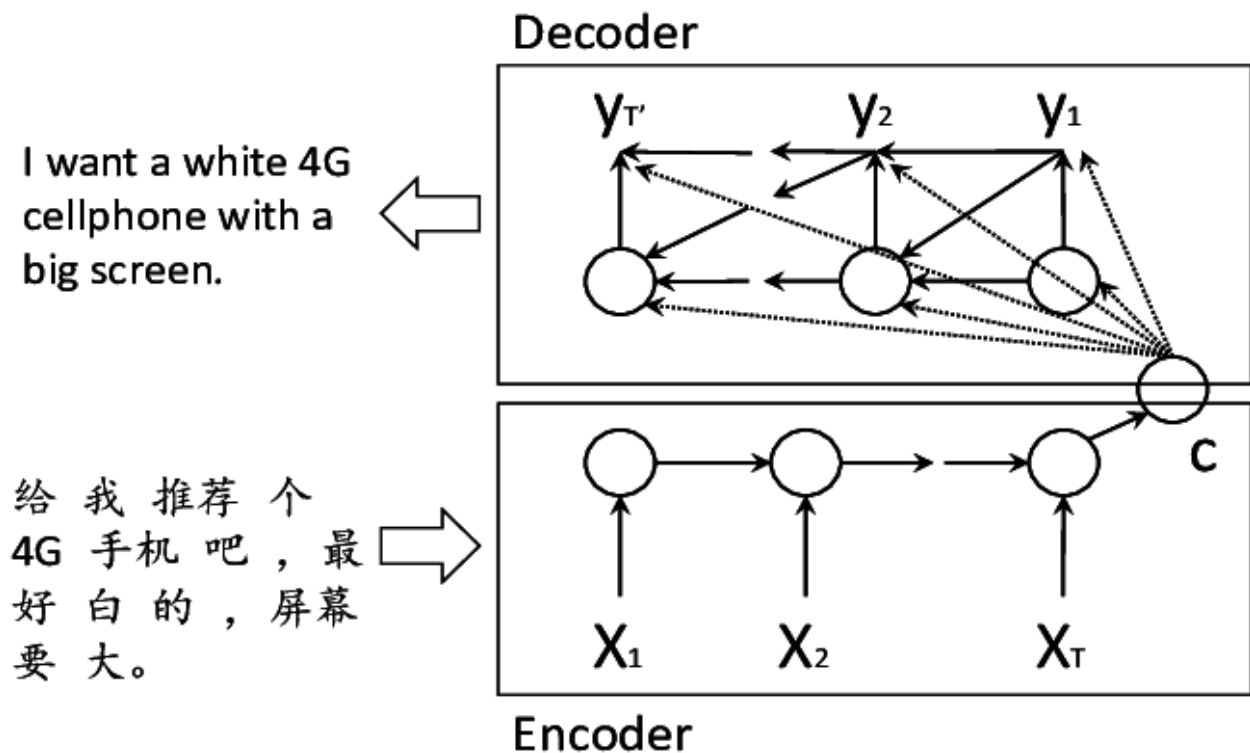


Fig. 2. An illustration of an RNN-based neural network model for Chinese-to-English machine translation.²

However, NMT has its flaws. The Encoder-Decoder architecture has problems translating long sequences of text as well as rare words. In addition, NMT systems are known to be slow to train. It takes a lot of time and computational resources to train an NMT system on a large-scale translation dataset, and due to the large numbers of parameters used in NMT systems they are generally much slower than phrase-based systems. Finally, NMT systems sometimes do not translate all the parts of the input sentence. (Wu et. al., 2016: 1-2) One of the newest NMT systems which aims to solve

² Source: Wu, Yonghui et al. 2016. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*

the aforementioned problems is GNMT, an NMT system produced by Google to improve fluency and accuracy in Google Translate, Google’s proprietary translation software. According to Wu et. al. (2016: 1), GNMT “reduces translation errors by an average of 60% compared to Google’s phrase-based production system.”

In addition to Google Translate, many other translation tools exist. Some are new and some have been around for decades. This includes Systran, Reverso, Transcend, E-lingo, Alphaworks and Compreno, among others.

Translator	Address	Characteristics
Alphaworks®	http://www.alphaworks.ibm.com/aw.nsf/html/mt	Translates English into seven languages; transfer method
Compreno®	http://www.abbyy-developers.eu	Meaning-based text analysis
E-lingo®	http://www.elingo.com/text/index/html	Twenty pairs of languages available; transfer method
Reverso®	http://trans.voila.fr	Thirteen pairs of languages available; transfer method
Systran®	http://www.systransoft.com	Twelve pairs of languages available; direct transfer method
Transcend®	http://www.freetranslation.com/	Eight pairs of languages available; direct transfer method

Fig. 3. Existing tools and software for MT

4. Advantages and disadvantages of machine translation

Machine translation would not have become as widespread as it now is if it did not offer some advantages. One of the biggest advantages of machine translation is speed; a computer is capable to producing a translation in seconds. Machine translation is also much more cost-efficient for translating large volumes of text. A variety of machine translation programs like the ones mentioned earlier offer a lot of flexibility in that regard. In short, machine translation offers a low-cost, quick and easy solution for translating a large amount of text of a non-literary nature..

On the other hand, machine translation's difficulties with translating literary texts are attributed to its inability to understand and properly translate certain local phrases and slang as well as complex terms of an industry-specific nature and to account for context when translating. Machine translation programs also do not perform many translation procedures employed by human translators: sentence splitting, shifts of function and/or category, explicitation³, modulation and paraphrasing. In addition, the length, information flow and structure of machine translated texts are more similar to the source text than to a text translated by a human. (Ahrenberg, 2017: 7) This can create a text that feels odd and robotic, one that cannot be accepted as is, and which must be revised by a human.

³ Explicitation is making something that is implicit in the source text explicit in the target text or making something that is already explicit in the source text even more explicit in the target text. (Kraskowska, 2018)

5. Computer-assisted translation

While similar to machine translation in that it is used to speed up the translation process, computer-assisted translation differs from machine translation greatly. According to Craciunescu et. al. (2004: 7), “computer-assisted translation is a complex process involving specific tools and technology adaptable to the needs of the translator.” In a translation process, the computer becomes a source of various tools which speed up the translation process or otherwise make it easier. Computer-assisted translation (CAT) tools include: electronic dictionaries, glossaries and terminology databases, online bilingual texts, concordances and translation memories. Consulting a dictionary on the computer, while ostensibly no different from consulting a paper dictionary, is much quicker as one only needs to type in the term they are looking for instead of leafing through a physical book. Bi-texts, electronically stored bilingual corpora consisting of a source text and its translation, are sources of pre-translated fixed expressions which also speed up the translation process. Computer concordances are “word-processing programs that produce a list of all the occurrences of a string of letters within a defined corpus with the objective of establishing patterns that are otherwise not clear.” (Craciunescu et al., 2004: 8) They are especially valuable for translating specialized texts with fixed expressions and vocabulary since they ensure that the translator maintains consistency in translating certain terminology. They are less useful for translating literary texts because literary texts make frequent use of metaphors and are characterized by polysemy. Translation memories are databases containing bilingual pairs that a translator previously encountered. If a translator encounters the word or phrase that they have encountered

before, the translation memory program will retrieve their previous translation and recommend it for the new translation. The translator can then accept the recommended translation, edit it to suit the new context or reject it entirely. (Craciunescu et. al. 2004: 7-9)

6. Method

The first step of this research was to compile corpora of texts for the analysis. The corpora include:

1. A corpus of original texts (i.e. source texts, ST) in Croatian,
2. A corpus of machine translated texts from ST Croatian into target texts (TT) in English,
3. A corpus of English TT translated by a human translator, as a control corpus.

The source texts as well as the human-translated ones all originate from the website for the Republic of Croatia's Ministry of Science and Education (mzo.gov.hr) and include a collection of documents. The texts are administrative in nature and include the government's recommendations on organizing the working day of pupils in the distance teaching and learning which was implemented in Croatia in 2020 during the COVID-19 epidemic, guidelines for assessment and grading in a virtual environment, guidelines for distance learning for primary and secondary schools, as well as several news articles concerning meetings and conferences attended by Croatian politicians which are also written in administrative style. Administrative texts are written in a way so as to be as understandable as possible, with an emphasis on clarity, accuracy, simplicity and

conciseness. This makes the texts relatively easy for machine translation systems to translate, unlike literary texts or texts written in everyday language. Such texts use many of colloquialisms, which makes texts written in everyday language the most difficult for machine translation systems to translate properly (Ljubas, 2017: 37.) The machine translated texts were translated by Google Translate using onlinedoctranslator.com, where I uploaded the entire corpus in the form of a single document.

7. Analysis

The corpora were analyzed in three steps.

7.1. Statistics

In the first step of the analysis, I used Sketch Engine to compare the differences between the different corpora based on the number of words and parts of speech (POS) present in each corpus. I did this in order to get an idea of how the corpora differ from each other on a basic level.

	MT texts	HT texts	Source texts
Total words	16,357	16,553	15,247
Nouns	3,959	4,210	4,424
Verbs	2,168	2,262	1,640
Adjectives	1,015	1,103	1,414
Adverbs	498	450	310
Pronouns	551	527	979
Conjunctions	931	877	1,473
Prepositions	2,558	2,592	1,782
Numeral	107	144	173

Table 1. Differences in number of words and POS between corpora

As can be seen in Table 1, there are some differences between the corpora. Both of the translations – both the machine translated and human translated ones – have more total words compared to the ST. The ST also features significantly more pronouns, conjunctions and adjectives but fewer verbs and prepositions. Curiously, there are no particularly striking differences between the machine translated and human translated texts, perhaps aside from the fact that the machine translated texts have slightly fewer nouns.

7.2. Quantitative analysis

In the beginning of my quantitative analysis of the MT corpus, I used ProWritingAid and Grammarly, two extensions for Microsoft Word to determine how readable the MT texts were.

ProWritingAid's summary report gave the corpus an overall score of 66/100, grammar 74/100, spelling 80/100 and style 44/100. The three largest issues it uncovered include a low readability measure, which indicates that the text is difficult to read, a high "glue index", which indicates a large number of filler words in the text as well as many long sentences which make reading the text difficult – there are 203 long sentences in the text out of 797 sentences in total. To expand on the readability feature, ProWritingAid identifies the paragraphs in the text and divides them into three categories:

1. Easy-to-Read Paragraphs
2. Slightly Difficult-to-Read Paragraphs and
3. Very Difficult-to-Read Paragraphs.

It should be noted that ProWritingAid did not do a perfect job in identifying all of the paragraphs in the corpus. It overlooked many paragraphs in the process of analyzing the text. In total, it found 137 very difficult-to-read paragraphs, 27 slightly difficult-to-read paragraphs, and 8 easy-to-read paragraphs, despite there being a total of 412 paragraphs in the entire corpus, according to ProWritingAid itself.

The corpus garnered a score of 38,5 on the Flesch Reading Ease test, 13.3 on the Flesch-Kincaid grade level, 13.3 on the Coleman-Liau readability index, 13.9 on the Automated Readability Index and a 7-8 on the Dale-Chall readability formula. These

results are conflicting; a 13.3 on the Flesch-Kincaid grade level is indicative of a very difficult text with an estimated reading grade of a college graduate, while a 13.3 on the Coleman-Liau readability formula is typical for college student, as is a 38,5 on the Flesch Reading Ease test. Meanwhile, a 7-8 score on the Dale-Chall readability formula is indicative of a text which would be easily understood by an average 9th or 10th grade student and a 13.9 on the Automated Readability Index is typical for an eighth grader.

ProWritingAid recognized 2376 unique words from a total of 1829 word families in the corpus. Of those words, the most commonly used words include:

1. the (1184)
2. and (754)
3. of (649)
4. to (497)
5. in (482)
6. a (294)
7. is (278)
8. for (265)
9. students (203)
10. learning (192)

These results are not surprising, as all but the last two words on this list are some of the most commonly used words in English. Since a majority of the corpus surrounds various recommendations and guidelines for distance learning, the inclusion of the words “students” and “learning” is also unsurprising.

An element of the corpus that ProWritingAid recognized as being egregiously bad is style, with a passive index of 34, 87 hidden verbs and 31 repeated sentence starts.

I used Grammarly to see if it would be able to offer a different perspective on the text or identify problems with it that ProWritingAid did not. Grammarly analyzes the text and identifies “issues” in it, separating it into four categories: correctness, clarity, engagement and delivery. Correctness relates to spelling, grammar and punctuation errors, but Grammarly was underwhelming in this regard, missing many egregious errors in the text and frequently identifying correct sentences as errors.

However time-consuming it is, it is limited and challenges teachers and teachers to be able to complete the tasks that the "average student" needs to complete within a given time frame.

For example, in this sentence Grammarly mistakenly advises to put a comma after the word “however”. In this case, “however” means “no matter how”. Because of that, adding a comma would create a nonsensical sentence.

The “clarity” category brings attention to issues that help make a text easier to understand. This mostly concerns issues of conciseness i.e. rewriting sentences which are wordy but otherwise correct or changing a sentence which is in the passive voice into active. Grammarly recognized 316 clarity issues in the corpus, making the corpus as a whole “a bit unclear”. The category “engagement” points out frequently overused words or words that appear repeatedly in the text. 224 engagement issues are present in the corpus, making engagement “a bit bland”. Only 5 issues of delivery were recognized,

making delivery just right. Four of these issues relate to a *to-infinitive* verb being split by a modifier and one issue relates to an unnecessary comma in a situation where the word “so” begins a subordinate clause, with Grammarly’s recommendation being to replace “so” with “so that” and removing the comma.

7.3. Qualitative analysis

The machine translated texts were analyzed for errors using a modified version of Simeon’s (2008: 108-109) classification of errors. Simeon classifies errors into eight separate categories:

1. untranslated words (words which remained in the source language)
2. omitted words
3. inserted words (words from the target language which were unjustifiably inserted)
4. lexical errors (words which significantly semantically differ from words in the source language)
5. spelling errors
6. morphosyntactic errors (incorrect forms, disagreement between subject and predicate, errors in clauses, wrongly chosen functional words etc.)
7. stylistic errors (awkwardly formulated syntagma, translation equivalents which are semantically similar to the source word, but which may lead to misunderstanding or which may confuse the reader)
8. word order errors

For the purposes of this paper, I have effectively removed word order errors, opting instead to consider word order errors morphosyntactic errors. After my initial analysis of the corpus, I also decided to add two more error categories: punctuation errors and inconsistencies.

It should be noted that errors that appeared several times were not counted more than once. With that being said, the final error count in the corpus is as follows:

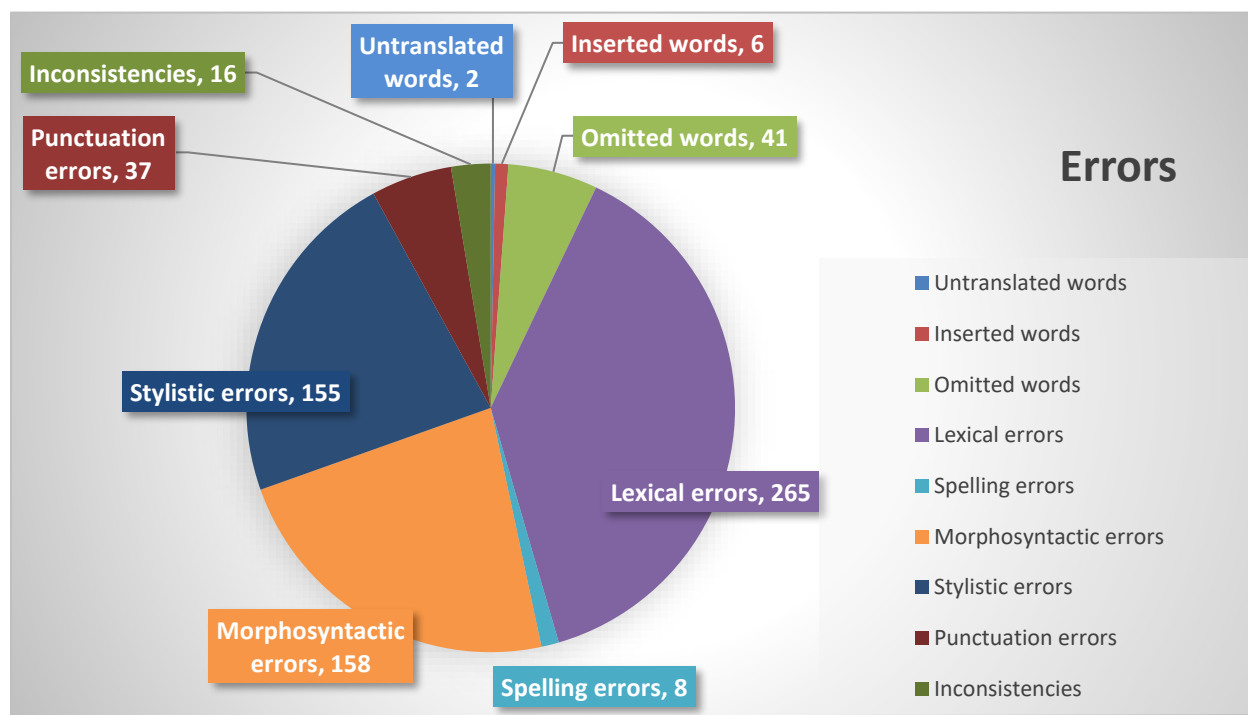


Fig. 4. Distribution of errors in the MT corpus

As can be inferred from Figure 1, lexical errors, morphosyntactic errors and stylistic errors dominate the corpus, making up 89% of all of the errors. I will now analyze each category of error one by one. As I use examples, the words and phrases I will be commenting on will be in **bold**. I will not be paying any attention to any other errors in a given sentence which are not in bold. The original source text will also be shown.

7.3.1. Untranslated words

Only two cases of untranslated words are present in the corpus.

Example 1	Učenici prisustvuju Školi na Trećem Čitaju, izražavaju se u pisanome i govornome obliku Računaju, rješavaju matematičke zadatke, modeliraju Vježbaju (TZK)
	Students attend the School at Third They read, express themselves in written and spoken form They calculate, solve mathematical problems, model Exercise (TZK)

In this example, Google Translate failed to translate TZK as PE or physical education. While this could imply a difficulty of MT systems to translate acronyms, and this coincides with other occurrences of acronyms in the corpus (see 7.3.4.), not enough acronyms are present in the corpus for it to be possible to draw conclusions. It is also possible that the term TZK was not translated because it was not found in the context of a sentence, being instead found in an example of a schedule for students in the form of a list.

Example 2	Obavijest o odgodi HR PRES konferencije „Učitelji i nastavnici budućnosti”
-----------	---

Notice of postponement of HR **PRES** conference “Teachers and teachers of the future”

The term “pres konferencija” is a loanword, an anglicism which has a proper Croatian form – novinska konferencija. Nevertheless, the term was translated as “pres conference”, suggesting that Google Translate does not recognize the anglicism.

7.3.2. Omitted words

There is a total of 41 omitted words, phrases, parts of sentences or entire sentences in the corpus. A recurring error throughout the text is an incomplete sentence:

Example 1 Pri tome su učenici i nastavnici ključni jer trebaju materijalima koji su već dostupni u nekome od centralnih servisa pridružiti i vlastite materijale, **upute i pratiti učenika, ali da sve to bude u razumnim i prihvatljivim vremenskim okvirima te prilagođeno spoznajnim mogućnostima učenika.**

Students and teachers are crucial in that they need to attach their own materials to materials already available in one of the central services,

Bizarrely, errors like this occur with frequency in the corpus; 8 cases of omitted words in the corpus are such incomplete sentences, while another 3 cases are sentences which have been completely omitted.

Example 2 Međutim, ako se nastavnici odluče ne oslanjati se na lekcije koje su pripremljene u centralnim servisima, onda slične materijale trebaju pripremati samostalno i davati jasne upute učenicima u svim izvorima koje trebaju koristiti. **Napominjemo da se isključivo izvođenje nastave na daljinu putem videokonferencije ili audiokonferencije u realnome vremenu na kojima sudjeluju svi učenici ne preporuča zbog velikog opterećenja online sustava, ali i zato što postoje primjereniji načini organizacije nastave na daljinu.**

However, if teachers choose not to rely on lessons prepared in central services, then similar materials should be prepared on their own and provide clear guidance to students in all sources to use.

Every such example of an omitted sentence or part of sentence occurs at the end of a paragraph. It is not easy to tell why this error occurs; it is possible that this is an error of onlinedoctranslator.com, seeing as Google Translate was able to produce a translation when that same paragraph was pasted into Google Translate directly.

Example 3

Učenici, s druge strane, ističu da najbolje uče kad uživaju u onome što uče i kako uče, i tad **im** vrijeme brzo prođe.

Students, on the other hand, point out that they learn best when they enjoy what they are learning and how they are learning, and then time passes quickly.

Sometimes omitting a word fails to produce a noticeably difficult to understand translation; in fact, it only slightly changes the meaning of a sentence. In Example 3, it would be most correct to say “and then time passes quickly *for them*”. However, this is not always the case.

Example 4

Možda se čini da **je** 5 sati dnevno **zahtjevno**, ali ne smije se zaboraviti da su do uvođenja nastave na daljinu učenici u fizičkoj učionici svaki dan provodili najmanje 4 sata, a poslije su učili kod kuće ili u produženome boravku te da su još i vikendom imali neke obveze oko nastave.

It may seem like 5 hours a day, but it should not be forgotten that until the introduction of distance learning, students spent at least 4 hours each day in the physical classroom, then later they studied at home or in extended residence and even had weekends some teaching responsibilities.

In this example, omitting “je zahtjevno” produces a nonsensical translation.

7.3.3. Inserted words

6 cases of inserted words, phrases or sentences were found in the corpus.

Example 1 Ogledni primjer rasporeda za učenike razredne nastave tijekom radnoga tjedna (ponedjeljak – petak):

An **example** example of a work week schedule (Monday – Friday) for elementary school students:

A recurring error by Google Translate, I counted “example example” as a case of both an inserted word and a lexical error.

Example 2 Promjena paradigme učenja i poučavanja koja se očituje u pomaku s činjeničnog znanja na razvoj kompetencija nužnih za uspješan život i rad u 21. stoljeću, kao što su kritičko razmišljanje, rješavanje problema i informirano donošenje odluka, razvoj kreativnosti i inovativnosti, mora biti vidljiva i u metodama vrednovanja.

MT The shift in the learning and teaching paradigm **shifting** from factual knowledge to the development of competencies necessary for successful life and work in the 21st century, such as critical thinking, problem solving and informed decision making, development of creativity and innovation, must be visible in methods as well evaluation.

HT The change of the learning and teaching paradigm, from teaching factual knowledge to developing competences necessary to live and work in the 21st century, such as problem solving and informed decision making, has to be manifest in the assessment methods.

In this example, the word “shifting” is redundant as it is mentioned at the beginning of the sentence.

7.3.4. Lexical errors

Lexical errors are any errors where something is translated using a word or phrase with a denotative or connotative meaning that does not match the denotative or connotative meaning of the original. Lexical errors were by far the most common error type in the corpus, with a total of 265 cases. The most frequent lexical errors are simply mistranslated words, including the first sentence of the corpus:

Example 1 Izvođenje nastave na daljinu predstavlja izazov za učitelja i učenika, ali i za cjelokupnu obitelj jer se proces učenja i poučavanja odvija kod kuće pa je potrebno osigurati određene preuvjete za učenje, ali i procijeniti koliko zapravo treba učiti.

MT **Teaching** distance learning is a challenge for both the teacher and the student, as well as for the whole family because the learning and

teaching process takes place at home, so it is necessary to provide certain prerequisites for learning, but also to evaluate how much one actually needs to learn.

HT Implementing distance learning is a challenge to both the teacher and the student, but also for the entire family because the process of learning and teaching is taking place in the home where it is necessary to ensure certain conditions for studying.

These errors are sometimes not severe enough to significantly impact the reader's ability to understand the sentence, as in Example 1. They still, however, convey a different meaning than the original. Other times, the translation is so erroneous that it is difficult to understand what is being said or the meaning of the translation is completely incorrect.

Example 2 Primjerice, većina učenika ne može naučiti razlomke samostalno iz udžbenika.

MT For example, most students cannot learn **the passages** independently from the textbook.

HT For example, most students cannot learn fractions by themselves by just using the textbook.

Sometimes the difference between the correct translation and the MT is subtle:

Example 3 Učitelji šalju dodatne zadatke vezane uz planirane teme roditeljima, ali pri tome trebaju voditi računa o opterećenju učenika i svrsishodnosti zadataka, kao i tehničkim mogućnostima koje učenici imaju kod kuće.

MT Teachers send additional tasks related to the planned topics to parents, but in doing so, they should take into account the student workload and **purposefulness** of the tasks as well as the technical capabilities that students have at home.

HT Teachers send additional tasks connected to planned topics to the parents, but while doing so they need to take into consideration the students' workload and the purpose of the assignments, as well as technical conditions in the home.

In this example, “svrsishodnost” would be correctly translated as “purpose”, as purposefulness is the quality of having purpose.

Curiously, a common error in the corpus are the various different translations of the acronym MZO (Ministarstvo znanosti i obrazovanja):

Example 4 It is a priority to ensure that accurate and complete information is flowed between the **MoES** and school principals, between school principals and educational staff, and between school and students and parents.

If the teacher has finished his / her classes, his / her students do not have to follow the content provided by the **MHE**, but the subject teacher in the virtual class delivers the content and activities, but then it must be clearly stated on the school web pages.

Please note that there are subjects that are not covered by the central schedule published by the **WHO**, and teachers themselves should prepare materials and instructions for such subjects.

Secondary schools that enrolled students whose families are guaranteed minimum benefits have received funding from the **MoH** for the purchase of tablets for students with lower socioeconomic status, or for whom the school estimates they do not have a computer or internet access at home.

Lexical errors also included any translations of names of documents or other names that already have official translations e.g. *Preporuke o organizaciji radnog dana učenika tijekom održavanja nastave na daljinu* has an official translation - Recommendations for organizing a student's workday in distance teaching and learning. Therefore, the MT of "Recommendations on organizing students' work days during distance learning" was not accepted.

A particular problem for Google Translate was translating the terms *vrednovanje za učenje, vrednovanje kao učenje* and *vrednovanje naučenog*. Because these terms have a translation equivalent in the form of *assessment for learning, assessment as*

learning and assessment of the learned, the translation “evaluation for learning, evaluation as learning and evaluation of the learned” was considered a lexical error.

On a similar note, *vrednovanje i ocjenjivanje* was translated as many as four different ways, with three of those being lexically incorrect (see 7.3.9.)

It should also be noted that in rare circumstances Google Translate was not able to understand slight nuances in meaning. When translating the word *kompetencije*, it translated it as “competences” instead of “competencies”.

7.3.5. Spelling errors

All cases of spelling errors in the corpus – all eight of them – are cases of incorrect capitalization.

Example 1 Ravnatelj iz Karlovca poslao nam je sljedeću poruku: “...Samo bih izvijestio o tome da se nastava na daljinu provodi s velikim (čak malo neočekivanim!) uspjehom.

MT The director from Karlovac sent us the following message: “... I would just like to report that distance learning is conducted with great (even a little unexpected!) **Success**.

HT A school principal from Karlovac sent us the following message: “I would just like to report that distance learning is a big (and a bit unexpected!) success.

In this example, it is likely that the MT capitalized the word *success* because there is an exclamation point before it. It is not surprising that MT would struggle with this considering the incorrect sentence format. Other cases of wrong capitalization are not as easy to explain:

Example 2 Sve ravnatelje molimo da prate obavijesti u učionici za ravnatelje na Loomenu.

MT We ask all **Principals** to follow classroom announcements for Loomen **Principals**.

HT We ask all school principals to pay close attention to notifications in the virtual classroom for principals on Loomen.

Example 3 Zaključno je razmotrena prijava San Marina za članstvo u Europskome prostoru visokog obrazovanja i Nacrt dnevnoga reda sastanka svih država koje sudjeluju u provedbi Bolonjskoga procesa koji će se održati u Kijevu, Ukrajina, 4. - 5. ožujka, uz supredsjedanje Hrvatske.

San Marino's application for membership of the European Higher Education Area and the Draft Agenda for the **Meeting of all States**

participating in the Bologna Process to be held in Kiev, Ukraine, March 4-5, co-chaired by Croatia, were discussed.

Example 3 highlights a common cause of incorrect capitalization: Google Translate often times assumes that names of documents, committees etc. are longer than they are, and so capitalizes parts that don't need capitalization. In this case, *Nacrt dnevnog reda* was translated as "Draft Agenda for the Meeting of all States" instead of just "Draft Agenda".

7.3.6. Morphosyntactic errors

Morphosyntactic errors are some of the more common error types in the corpus. There are 158 morphosyntactic errors in total. In addition to being some of the most common errors, there is a large variety of morphosyntactic errors in the text. Morphosyntactic errors include word order errors, subject-verb disagreement, incorrect word structure, wrongly chosen function words, missing function words, missing articles, missing determiners, wrong determiners or any other errors in the structure of words or sentences, such as connecting independent clauses by a comma or changing the subject of a clause.

Example 1 Što se uvjeta za učenje tiče, to podrazumijeva opremu za nastavu za daljinu (upute o tome dane su školama) i potrebne udžbenike, ali i da se za učenike kod kuće organizira mirni kutak za učenje.

As far as the conditions of learning are concerned, **this includes** the equipment for distance learning (instructions are given to schools) and the necessary textbooks, but also **to organize** a quiet corner for students at home.

Subject/verb disagreement was a relatively rare error, appearing only several times in the corpus.

Example 2 Vrijeme provedeno u učenju podrazumijeva vrijeme u kojem se poučava (u školi, odnosno putem sustava za nastavu na daljinu), ali i vrijeme samostalnoga ili suradničkoga učenja. Takvo ukupno vrijeme ovisi o uzrastu djeteta i kreće se od 5 do 8 sati dnevno tijekom radnoga tjedna.

The time spent **in** learning includes the time in which it is taught (at school, or through the distance teaching system), but also the time **of** independent or collaborative learning.

Example 2 shows two morphosyntactic errors: the first is an unnecessary *in*, as the word spend does not require a preposition. The second error is one of the more common morphosyntactic errors in the corpus: a wrongly chosen function word. When a function word is translated incorrectly by MT systems, it is often translated literally. Other examples include:

Example 3

Možda se čini da je 5 sati dnevno zahtjevno, ali ne smije se zaboraviti da su do uvođenja nastave na daljinu učenici u fizičkoj učionici svaki dan provodili najmanje 4 sata, a poslije su učili kod kuće ili u produženome boravku te da su još i vikendom imali neke obveze oko nastave.

It may seem like 5 hours a day, but it should not be forgotten that **until** the introduction of distance learning, students spent at least 4 hours each day in the physical classroom, then later they studied at home or in extended residence and even had weekends some teaching responsibilities.

Example 4

Neki od tih kolokvija mogu se zamijeniti rješavanjem zadataka koji se preuzimaju iz baze zadataka, tako svaki student dobiva svoje zadatke.

Some of these exams can be replaced **by** solving assignments that are downloaded from **the** assignment database, so each student gets his or her assignments.

In Example 3, a better choice of function word would have been *before*, highlighting MT systems' limited ability to understand the context of a sentence. Example 4 highlights the same problem, as well as another common error in the corpus: incorrect articles – seeing as this is the first mention of an assignment database, the article *an* would have been correct.

Example 5 Učenici, s druge strane, ističu da najbolje uče kad uživaju u onome što uče i kako uče, i tad im vrijeme brzo prođe.

Students, on the other hand, point out that they learn best when they enjoy what they are learning and how they are learning, **and** then time passes quickly.

In this example, the function word *and* is incorrectly used as this is a result clause; a better choice would have been *so that*.

Example 6 Promjena paradigme učenja i poučavanja koja se očituje u pomaku s činjeničnog znanja na razvoj kompetencija nužnih za uspješan život i rad u 21. stoljeću, kao što su kritičko razmišljanje, rješavanje problema i informirano donošenje odluka, razvoj kreativnosti i inovativnosti, mora biti vidljiva i u metodama vrednovanja.

The shift in the learning and teaching paradigm shifting from factual knowledge to the development of competencies necessary **for successful life** and work in the 21st century, such as critical thinking, problem solving and informed decision making, development of creativity and innovation, must be visible in methods as well evaluation.

Missing determiners and articles were a common error in the corpus.

Example 7 Nastava na daljinu odvija se u izvanrednim okolnostima i treba iskoristiti sve njezine prednosti, a pokušati umanjiti većinu nedostatka.

Distance learning takes place in extraordinary circumstances and **should take full advantage of it**, and **try to reduce most of the disadvantage**.

Verb structure errors were uncommon, but the sentences marred by such errors were difficult to understand.

Example 8 Načela vrednovanja u nastavi na daljinu

Valuation principles in distance education

Example 8 shows a noun structure error, a one-of-a-kind error in the text.

Example 9 Kontakt učitelja, nastavnika s učenicima i roditeljima važan je kako bi se stvorio krug povjerenja za učenje, u tome svatko ima svoju ulogu.

Contacting teachers, teachers with students and parents is important in order to create a circle of learning confidence, in which everyone has a role to play.

Example 9 presents a one-of-a-kind error in the corpus: change of subject.

Example 10

U suradnji sa stručnim suradnicima u školi potrebno je nastaviti provoditi i u ovim okolnostima upute za prilagodbu za učenike s teškoćama, i to u skladu s rješenjima o prilagodbi i individualizaciji nastavnih sadržaja, a za darovite učenike osmisliti zadatke koji su zahtjevniji i kojima se propituju više kognitivne razine.

In conjunction with the school's professional associates, it is necessary to continue implementing the adjustment instructions for students with disabilities in these circumstances, in accordance with decisions on the adaptation and individualization of teaching content, **and for gifted students to design tasks that are more demanding and more cognitive. levels.**

In cases of particularly long sentences, the structure of the sentence sometimes completely fell apart, leaving them disjointed and impossible to understand.

Example 11

Aktivnost i samostalnost u radu učenika treba prepoznati, poticati i pozitivno vrednovati, počevši od domaćih zadaća preko suradničkog učenja do samostalnog istraživanja.

In particular, in all subjects, student activity **through discussions and homework can be evaluated** and at least one grade added.

Not many word order errors were found in the corpus, and it bears mentioning that word order errors were more common in sentences which were either very long or strangely worded even in Croatian:

Example 12 Konkretno, u svim predmetima može se vrednovati aktivnost učenika preko diskusija i domaćih zadaća i tome pridodati barem jedna ocjena.

In particular, in all subjects, student activity **through discussions and homework can be evaluated** and at least one grade added.

A unique morphological error was change of perspective/voice:

Example 13 Učitelji trebaju voditi računa o opterećenju učenika. To podrazumijeva da se trebaju postići planirani odgojno-obrazovni ishodi, ali i da u sadržajima treba razlučiti ono što je bitno od onoga što je sporedno ili nije nužno

Teachers should take into account the workload of students. This implies that the planned educational outcomes should be achieved, but **also that the contents should distinguish** what is important from what is secondary or not necessary.

According to the above example, it is the contents themselves which should distinguish what is important and not the teachers, which is incorrect.

7.3.7. Stylistic errors

Stylistic errors include any awkwardly formulated syntagma or translation equivalents which are semantically similar to the source word, but which may lead to misunderstanding or which may confuse the reader. An obvious flaw of this error classification is that it is somewhat open to interpretation; I needed to rely on my personal judgement to determine whether a sentence was formulated awkwardly enough for it to be considered a stylistic error or not. That being said, I found stylistic errors to be one of the most common types of errors in the corpus.

Example 1 Posebni je izazov vrednovanje praktičnih vještina, onih koji se nalaze u psihomotoričkoj domeni, samo neke od njih mogu se djelomično demonstrirati i u virtualnom okruženju.

MT **The particular challenge is to evaluate practical skills**, those in the psychomotor domain, only some of which can be demonstrated in a virtual environment.

HT Practical skills are particularly challenging to assess, especially psychomotor skills which can be demonstrated in a virtual environment only to an extent.

This is one of the most representative examples of a stylistic error found in the corpus. You can see how the human translation differs from the MT. The MT sentence

has the same structure as the original, while a human changed the structure of the sentence to better fit in English by changing the subject of the sentence. This is one of the biggest drawbacks of MT: it is not very flexible. Many stylistic errors can be attributed to this.

Example 2 Dakako, izvrsno je i da drugi učenici mogu odgovoriti na postavljena pitanja i probleme, što je poželjan scenarij.

MT Of course, **it's** great that other students can answer the questions and problems they ask, which is a desirable scenario.

HT In addition, other students can give answers to the questions posed and provide solutions to problems, which is a desirable scenario.

Considering the formal nature of the text, abbreviations are inappropriate which is why I considered this to be a stylistic error. The differences in the way that MT and HT phrase sentences can also be seen in this example. The original never specified who poses the questions and problems, so the HT is stylistically more appropriate in that regard as well.

A frequently repeated stylistic error was repeated words: *problemski zadatak* translated as “problem problem”, *učitelji i nastavnici* translated as “teachers and teachers”, *učenici i student* as “students and students”.

7.3.8. Punctuation errors

I implemented punctuation errors as an error category when I noticed the plethora of punctuation errors in the corpus. I deliberated over considering a missing Oxford comma an error or not but ultimately decided not to. Despite being obligatory in British English, it is optional in American English, and there is never a moment in the corpus when a lack of an Oxford comma leads to misunderstanding. There are, however, many other punctuation errors in the corpus – a whole 37 of them.

Example 1 Pri tome učenici ne trebaju gledati TV program po rasporedu ako nije usklađen s njihovim rasporedom, ali svakako nastavnici trebaju djecu uputiti na videolekcije, tj. trebaju pronaći odgovarajuće ako nisu izradili svoje

In doing so, students do not need to watch a TV program on a schedule if it is not in line with their schedule, but certainly teachers need to refer children to video lessons, **ie** they need to find the appropriate ones if they have not made their own.

Every time the Latin abbreviations “i.e.” and “e.g.” were used, they were written as “ie” and “eg”. I considered this to be an error as writing the abbreviations without full stops in between the characters is less formal, which is inappropriate for this style of text.

Example 2 Uz mogućnost slučajnog odabira pitanja, test je moguće konfigurirati tako da i odgovori u zadacima višestrukog izbora unutar pitanja budu drugačijeg redoslijeda.

MT With the possibility of randomly selecting questions, the test can be configured so that the answers in the **multiple choice** tasks within the question are in a different order.

HT In multiple choice questions, the answer sequence can be randomized as well as the order of appearance of questions.

Missing hyphens were relatively rare. Other words that were missing hyphens include: “third-year”, “e-learning”, “time-consuming” and “back-up”. Curiously, this was an error made even by the human translator.

Example 3 Učitelji se u ovakvoj situaciji nikad nisu našli i teško je brzo promijeniti način poučavanja i model izvođenja nastave, a još je teže promijeniti metode vrednovanja.

MT Teachers have never found themselves in this situation **and** it is difficult to quickly change the teaching method and model of teaching, and it is even more difficult to change the methods of evaluation

HT The current situation is novel for all teachers and it is difficult enough to change one’s teaching methods, let alone assessment methods.

In this situation, a comma is missing before a compound sentence. Interestingly, the human translator again made the same error as the computer, not including a comma themselves.

Example 4 O tome je napisano dosta literature, iako je ona pretežito usmjerena na tercijarno obrazovanje (više i visoko), a manje na predtercijarno obrazovanje (predškolsko, osnovnoškolsko i srednjoškolsko), može se primijeniti na sve razine obrazovanja.

MT Much literature has been written about it, **although** it is mainly focused on tertiary education (higher and higher) and less on pre-tertiary education (preschool, primary and secondary), it can be applied to all levels of education.

HT This is shown by substantial research, which, even though done primarily in tertiary education, and to a smaller degree in pre-tertiary education, is applicable at all levels of education.

This is an example of the opposite problem, two independent clauses joined by a comma.

Example 5 Međutim, opet naglasak ne treba biti na samom konačnom proizvodu nego procesu učenja, i to tako da učenik treba najprije predati nacrt rada ili plan istraživanja za koji dobiva povratnu informaciju vršnjaka, a onda i učitelja.

MT However, again the emphasis should not be on the final product itself but on the learning process, **so that** the student should first submit a work plan or research plan for which he or she receives peer feedback and then the teacher.

HT Again, the emphasis should not be on the result itself, but on the learning process, so that the student first needs to submit a draft or a research plan and collect feedback from peers and then from the teacher.

Example 5 shows a comma separating a dependent and independent clause.

Example 6 Do kraja tjedna ćemo sukladno rasporedu putem dostupnih aplikacija, Office Teams ili Google, zajedno prokomentirati dobivene rezultate putem chata ili sl.

By the end of the week, we will comment on the results obtained through the available applications, Office Teams or Google **together**. through chat or etc.

By the end of the week, we will use the tools available, Office Teams or Google chat, to comment on the results.

In this case, a full stop is incorrectly inserted mid-sentence. A surprisingly large number of such “floating full stops” can be found in the text.

7.3.9. Inconsistencies

Inconsistencies include any words or clauses which are translated in more than one way throughout the corpus. These include:

ST	MT
radni tjedan	work week workweek
vježbaju	exercise practice
odmore se	they rest it rests it breaks he rests
procjenjivanje i ocjenjivanje	assessment and assessment evaluation and evaluation evaluation and grading

	evaluation and assessment
Škola za život	School for Life School of Life
razredna nastava	class teaching classroom teaching
virtualni razred	Virtual Chamber virtual classroom
matematika, fizika i kemija	Mathematics, Physics and Chemistry mathematics, physics and chemistry
Agencija za strukovno obrazovanje i obrazovanje odraslih	Agency for Vocational Education and Training Agency for Vocational and Adult Education
Pravilnik o načinima, postupcima i elementima vrednovanja učenika u osnovnoj i srednjoj školi	Rulebook on Ways, Procedures and Elements of Student Evaluation in Primary and Secondary Schools Rulebook on Ways, Procedures and Elements of Student Assessment in Primary and Secondary Schools
razrednik	class teacher homeroom teacher

	classmate classroom teacher classroom
„Europa koja se razvija“	“Europe that is developing” “Europe in development”
Nacrt rimskog priopćenja	draft Communication from Rome Draft Roman Communication
vrednovanje kao učenje	assessment as learning evaluation as learning
vrednovanje naučenog	assessment of learned evaluating what is learned evaluating of the learned
MZO	MoH
	MoES WHO MHE

It is interesting to see the different ways that Google Translate translates the same words/phrases, but because consistency of translation is important, every such inconsistency had to be counted as an error.

8. Discussion

My results show that Google Translate is not very reliable in producing accurate and fluent translations compared to human translators. The litany of errors, particularly lexical and morphosyntactic errors negatively influence the reader's understanding of the text severely enough to be unacceptable. Many errors, particularly lexical and stylistic ones, can be attributed to Google Translate's inability to take into account context, particularly the context of the previous sentence. Due to machine translation systems' inability to account for context, polysemous and synonymous words are frequently mistranslated, e.g. "replay" instead of "rerun" when translating *repriza* or translating *oni posade biljku* as "they plant crew". In this example, Google Translate translated *posade* as crew (posada) instead of "to plant". This issue was not limited only to translating nouns, as is visible in the following example:

Također, mogu pokušati interpolirati funkciju u dane podatke te pritom iskoristiti prikladne matematičke alate kako bi se objasnila i modelirala neka stvarna situacija, npr. širenje lažnih vijesti ili virusa, rast stanovništva i sl.

I can also try to interpolate the function into the given data while using appropriate mathematical tools to explain and model a real situation, such as spreading false news or viruses, population growth, etc.

This is an example of grammatical polysemy – the computer has no way of knowing who the subject in the sentence is, as the only way for it to know it would be to read the context from the previous few sentences. That being said, the subject in this sentence is *they*, with *them* referring to students, making this a mistranslation.

Pitanja koja se postavljaju mogu uključivati povezivanje sadržaja s vlastitim iskustvima i pritom je središte vrednovanja na procesu učenja i svjesnosti o učenju, a ne na naučenim činjenicama.

The questions that may be asked may include linking the content to one's own experiences and, **at the same time**, it is the center of evaluation on the process of learning and awareness of learning rather than on learned facts.

In this example, *pritom* was translated as “at the same time”, which does not suit the context – another example of the problem of polysemy.

Interestingly, Google Translate made the error of translating Croatian names, so that *Stipe Župan* was translated as “Stipe Mayor”, another example of the issue of polysemy. However, more than that, it sometimes changed the names of people in unexpected ways, e.g. *Branka Ramljak* – “Branko Ramljak”; *Tome Antičić* – “Tom Anticic”. In addition to the obvious issue of changing *Branka* and *Tome* to “Branko” and “Tom”, it was inconsistent in the use of Croatian characters, sometimes translating names using the letters č, ć, and š, and sometimes not.

Word combinations like *učitelji i nastavnici* were common in the source text, and mistranslated as “teachers and teachers”. Unlike human translators, Google Translate never omitted the second “teachers” to produce a sentence that would fit more in the spirit of the English language. Other examples include: *učenici i studenti* – “students and students”, *dizajniranje i projektiranje* – “designing and designing”, *poster i plakat* – “poster and poster” etc., highlighting the issue Google Translate had with synonymy.

Google Translate also struggled with translating particularly long sentences. Long sentences often times ended up disjointed, with full stops inserted mid-sentence and parts of sentences missing. It bears mentioning that many cases of omitted ends of sentences occurred when paragraphs were very long as well.

Other issues include terms which do not have exact equivalents in English, such as *razredna nastava*, which relates to the first four years of elementary school, where a class of students is taught by a single teacher, as opposed to *predmetna nastava*, where a different teacher teaches every subject between year 5 and year 8 of elementary school. While *predmetna nastava* was given the adequate translation of “subject teaching”, *razredna nastava* was translated as either class teaching or classroom teaching.

It should be noted that Google Translate was sometimes able to translate entire paragraphs exceptionally well, which was surprising. An example which stands out is:

U Ministarstvu znanosti i obrazovanja održan sastanak Odbora
Skupine za praćenje provedbe Bolonjskoga procesa (Bologna
Follow-up Group Board)

Dana 7. veljače 2020. u sklopu hrvatskoga predsjedanja Vijećem EU-a održan je sastanak Odbora Skupine za praćenje provedbe Bolonjskoga procesa („Bologna Follow-up Group Board” - BFUG).

A meeting of the Bologna Follow-up Group Board was held at the Ministry of Science and Education

On 7 February 2020, a meeting of the Bologna Follow-up Group Board (BFUG) was held as part of the Croatian Presidency of the EU Council.

In this example, Google Translate dealt with parentheses and acronyms very well, using them correctly in place of full names. It is likely that the entire sentence is present in its memory.

Finally, it should be mentioned that sometimes, sentences were constructed in the ST so poorly or at the very least oddly that it is not surprising that Google Translate was not able to translate them properly:

Do konačnog rezultata može biti još jedan korak – izlaganje u virtualnom razredu da bi se vrednovalo napravljeno, ali i da bi učenik dobio podršku i uputu za daljnji rad.

Until the final result can be another step - presenting in a virtual classroom to evaluate what has been done, but also to give the student support and guidance for further work.

Ponovimo da je **pisani oblik rješavanja tipičnih zadataka**, kakvi se javljaju u Matematici i prirodoslovnim predmetima, uobičajen u našim školama i obično se u razredu pokušava spriječiti prepisivanje tako da učenici dobiju različite grupe testova koji se rješavaju u ograničenom vremenu.

Let us reiterate that writing a typical task, as they occur in Mathematics and Science, is common in our schools and usually tries to prevent transcripts in the classroom so that students receive different sets of tests that are solved for a limited time

Both of these examples show very strangely formulated syntax – the phrase “pisani oblik rješavanja tipičnih zadataka” is needlessly complicated.

9. Conclusion

When machine translation is discussed, people often voice concerns about it and its potential to replace human translators one day. While it is true that MT is getting better by the day, and even that it has gotten impressively good, the results of this study clearly show the limitations of such technology. With dozens of unacceptable errors on every page of text, it is difficult to imagine that machine translation systems like these could one day replace professional translators. This conclusion agrees with the findings of Schaiier (1996), who found that as many as 63 out of 69 machine translated texts translated using three different MT systems from English into Spanish were unacceptable in terms of correctness and comprehensibility. Another study by Çakır (2013) reached a similar conclusion for Google Translate translations from English into Turkish, with a 33.7% score for accuracy and 46.5% for fluency. It did, however, conclude, that the system helped translators produce successful translations faster by post-editing the MT instead of starting the translation from scratch. Ultimately, this kind of application is what machine translation was always designed for. The purpose of MT is not to produce perfect translations, it is to serve as aid for human translators in the production of translations, or at least to allow a person to understand the gist of a text. This, ultimately, I conclude Google Translate to be adequate for.

A limitation of this study is the fact that the errors were analyzed for essentially only one type of text – a text of an administrative nature. It would have been interesting to see how Google Translate tackled, for instance, a literary text, but that was unfortunately outside of the scope of this study. That, as well as analyzing errors produced

by Google Translate when translating from English to Croatian, would have allowed me to compare what kinds of errors the MT system makes in those situations.

I would like for research such as this to continue on the topic of machine translation, as machine translation technology could be a more beneficial tool in the translating profession, expediting the translation process if it were more reliable. Mirroring the recommendations of Ljubas (2017), there is a need for more concrete suggestions for improving MT systems to be constructed and for more money to be invested in their improvement. Further research could focus on the very limitations mentioned earlier, so that we could gain a more complete picture of where Croatian stands in terms of Google Translate's ability to translate to and from the language.

References:

- Ahrenberg, Lars. 2017. "Comparing Machine Translation and Human Translation: A Case Study." *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*: 21-28.
- Çakır, Sinan. 2013. "A study on the efficiency of "Google Translate" translation program." *Kırkıncı Yıl Yazıları*: 75-83.
- Craciunescu, Olivia et al. 2007. "Machine Translation and Computer-Assisted Translation: A New Way of Translating?" *Translation Journal* 8.
- DuPont, Quinn. 2018. "The Cryptological Origins of Machine Translation, from al-Kindi to Weaver." *Amodern 8: Translation-Machination*.
- Ljubas, Sandra. 2017. "Analiza pogrešaka u strojnim prijevodima sa švedskog na hrvatski." *Hieronymous 4*: 28-64.
- Hutchins, William John. 1995. "Machine translation: a brief history" in Koerner, E. F. K. et al., *Concise history of the language sciences: from the Sumerians to the cognitivists*. Cambridge: Elsevier Science, 431-445.
- Okpor, Margaret Dumebi. 2014. "Machine Translation Approaches: Issues and Challenges." *International Journal of Computer Science Issues* 11: 158-165.
- Schairer, Karen. 1996. "Can Computer Translation Replace Human Translation?" *Hispania* 79: 96-107.

Hutchins, John. 2004. "Two precursors of machine translation: Artsrouni and Trojanskij." *International Journal of Translation* 16: 11-31.

Web sources:

Brownlee, Jason. 2019. *A Gentle Introduction to Neural Machine Translation*. Webpage. <<https://machinelearningmastery.com/introduction-neural-machine-translation/>> Accessed 17th July 2020.

Kraskowska, Ewa. 2018. *Universals of Translation*. Webpage. < <https://bit.ly/32oNksy>> Accessed 24th August 2020.

Webpage. <mzo.gov.hr> Accessed 22nd July 2020.

Webpage. <onlinedoctranslator.com> Accessed 23rd July 2020.

Srivastava, Siddhant et. al. 2018. *Machine Translation: From Statistical to modern Deep-learning practices*. Webpage. < <https://arxiv.org/abs/1812.04238>> Accessed 20th July 2020.

Wu, Yonghui et al. 2016. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. Webpage. < <https://arxiv.org/abs/1609.08144>> Accessed 22nd July 2020.